# Instrumental rationality and proceeding acceptably over time

Chrisoula Andreou

*Abstract*: Theories of instrumental rationality often abstract away from the fact that actions are generally temporally extended and from crucial complications associated with this fact. Sergio Tenenbaum's *Rational Powers in Action* (2020) reveals and navigates these complications with great acuity, ultimately providing a powerful revisionary picture of instrumental rationality that highlights the extremely limited nature of the standard picture. Given that I share Tenenbaum's general concerns about the standard picture, my aim is to advance our general approach further by complicating and enriching debate regarding a picture of instrumental rationality that is accountable to the temporally extended nature of our actions and agency via the consideration of a few issues that merit further consideration and exploration. As I explain, despite stemming from or being associated with some important insights, some of the central ideas that Tenenbaum supports need to be qualified, modified, or reconsidered.

*Keywords*: cyclic preferences, incommensurability, instrumental rationality, satisficing, momentary versus extended actions, vague ends or projects.

Theories of instrumental rationality provide, roughly speaking, evaluations and imperatives regarding choice or action that figure as relative to certain basic given attitudes or stances of the agent. Such theories often abstract away from the fact that actions are generally temporally extended and from crucial complications associated with this fact. Sergio Tenenbaum's *Rational Powers in Action* (2020) reveals and navigates these complications with great acuity, ultimately providing a powerful revisionary picture of instrumental rationality that highlights the extremely limited nature of the standard picture (which focuses on the selection of momentary acts, chosen and effected – in auspicious cases wherein they are not blocked – at a choice point).[1] Given that I share Tenenbaum's general concerns about the standard picture, this symposium paper will lack the drama of a piece aimed at devastating criticism. Instead, my aim is to continue to advance the project of complicating and enriching

[1] All page references to Tenenbaum's work will be to (Tenenbaum 2020).

**philinq XI, 1-2023, pp. 99-108**
ISSN (print) 2281-8618-ETS doi: 10.4454/philinq.v11i1.444
Submitted: 14 November 2022
Accepted: 18 November 2022

debate regarding a picture of instrumental rationality that is accountable to the temporally extended nature of our actions and agency by raising some issues that merit further consideration and exploration.

My focus will be on three central ideas that Tenenbaum supports. First, I will focus on the idea that, given how the pursuit of ends over time often works, "someone may be irrational over a period of time without there being any moment during that time at which they were irrational" (viii). Second, I will focus on the idea that an instrumentally rational agent will often have to seek "acceptable" realizations of her ends rather than maximizing, and not due to the agent's bounded rationality but due to the nature of the ends themselves.[2] Finally, I will focus on the idea that an agent may be rationally permitted to waver between options in a way that involves her incurring costs that she could have avoided had she resisted "brute shuffling," though not if the costs are devastating.[3] As will become apparent, I think that each of these ideas needs to be either qualified, modified, or reconsidered, despite stemming from or being associated with some important insights.

With respect to the first idea, consider Tenenbaum's example of an agent with the vague and indeterminate end of writing a book. Suppose, in particular, that you are the agent in question and that, as Tenenbaum explains, the following conditions hold:

(i) [The project's] completion requires the successful execution of many momentary actions.
(ii) For each momentary action in which you execute the project, failure to execute that action would not have prevented you from writing the book.
(iii) On many occasions when you execute the project, there is something else that you would prefer to be doing, given how unlikely it is that executing the project at this time would make a difference to the success of your writing the book.
(iv) Had you failed to execute the project every time you would have preferred to be doing something else, you would not have written the book.
(v) You prefer executing the project at every momentary choice situation in which you could work on the project over not writing the book at all. (100-101)

For Tenenbaum, if, rather than succeeding, you failed to write the book as a result of having failed to execute the project every time you would have pre-

[2] Here and elsewhere, I use one of the singular personal (sometimes referred to as "preferred") pronouns "she," "he," or "they" rather than the unwieldy "she, or he, or they." I will not continue flagging instances in which "she, he, or they" is replaced with one of "she," "he," or "they."

[3] The phrase "brute shuffling" is borrowed from Michael Bratman (2012), who describes it as "lurching from one plan-like commitment to another incompatible commitment seen as equal or incomparable, in a way that involves abandoning one's prior intentions" (81).

ferred to be doing something else, you would count as irrational even though there is (Tenenbaum suggests) no particular moment at which you proceeded irrationally given that (by hypothesis) no particular momentary failure to pursue an end-directed action took you from being in a position to write the book to not being in a position to write the book.

Notably, Tenenbaum's view that some (rationally permissible) ends are indeterminate and vague is controversial, but I think he is right about this, and so I will accept this as common ground. Still, we should not jump to the conclusion that, in the contemplated case of failure, there is, other things equal, no moment at which you were (proceeding) irrational(ly). My reason for hesitation is based on the distinction between what is realized *in* a moment and what is being done *at* a moment (which I will briefly discuss here and which I say a great deal more about elsewhere).[4]

As Tenenbaum recognizes, doings are rarely momentary in the sense of being completed in a moment. Still, one can say of an agent engaged in the doing in progress of ϕ-ing between $t_1$ and $t_n$, that the agent is, *at*, say, $t_k$, ϕ-ing. For example, if the agent is making a cake between $t_1$ and $t_n$, then they are, *at*, say, $t_2$, when they turn on the oven, making a cake; importantly this holds even if they are interrupted and never complete the doing in progress of making a cake because they accidentally burn up the kitchen soon after turning on the oven. More generally, although a doing in progress *at* $t_x$ is not contained *in* $t_x$, and so we can say, to quote Michael Thompson (2008: 126), that the doing in progress "reach[es] beyond" $t_x$, its being in progress is not (to quote Thompson again) exposed to "simple disproof on the strength of what happens next" (2008: 126), since the doing in progress can be interrupted immediately after $t_x$.

To see that the distinction between what is realized *in* a moment and what is being done *at* that moment is potentially relevant in the above-mentioned failed book project case, consider the following: It may be that, although none of the agent's momentary doings – understood as doings completed in a moment – are irrational, the agent is, nonetheless, irrational *at* one or more of these moments because the agent is engaged in a doing in progress that reaches beyond her "momentary action" and is unacceptable relative to her end.[5] For instance, she may – given her dispositions and capacities, which, as Tenenbaum emphasizes, do not "easily show up in a snapshot of the agent's mind" (186) – be frittering away her life (which can be true at $t_x$ even if, unlike in the failed book project case of interest, her doing in progress of frittering away her life were, shortly after $t_x$, interrupted by, say, an unexpected transformation after a near death

[4] See, especially, (Andreou 2014), which I draw on in the next few paragraphs.

[5] For detailed discussion regarding relevantly similar cases, see (Andreou 2014).

experience prompting a life of great social and scholarly achievements). Perhaps there is invariably a problematic doing in progress in the cases of failure that are of interest.[6] Let me explain; and keep in mind that I am not assuming that there is anything inherently problematic about "frittering away" one's life (though there may be) but only that doing so is problematic if it is unacceptable relative to one or more of the agent's ends.

First note that I here allow, following Tenenbaum, that "one can be pursuing the end of ϕ-ing even while *at the same time* failing to take the necessary means to ϕ-ing, as long as pursuing an end extends through time" (128). As Tenenbaum explains, in such cases of failure, although one is failing to take the means to one's end, one is also doing certain things "that are intelligible only if taken as means to [one's] (failing) pursuit" (129). For example, in the book project case, one may be spending a great deal of time in front of one's computer with a Word document entitled "book" open (even if one also has several webpages open that one is browsing through). Note also that one need not be happy with the fact that one is, say, frittering away one's life to be accountable for this doing in progress. Relatedly, one can be accountable for this doing in progress just as one can be accountable for omissions like failing to take the necessary means to one's end; moreover, one can be failing to take the necessary means to one's end via the doing in progress of frittering away one's life.

Now, why not think that all cases of non-accidental failure (such as, for example, the case in which an agent, despite having the end of realizing long-term project P, has been frittering away her life, continues to fritter away her life, and ends her life having frittered it away) are ones in which, at least at some moments, there is a doing *in progress* that is incompatible with the agent's end, and that the agent is irrational *at* these moments, even if her momentary actions, which are contained *in* the relevant moments, are not irrational? This possibility should, I think, give us pause with respect to the suggestion that one may be irrational over a period of time without there being any moment during the relevant time frame *at* which one is irrational. Importantly, it can still be true that it is the irrationality of the doing in progress, which reaches beyond one's momentary actions, that explains one's irrationality at various moments during the relevant time frame rather than vice versa. And this point seems compatible with Tenenbaum's "non-supervenience thesis," according to which "the rationality of an agent through a time interval $t_1$ to $t_n$ does not supervene on the rationality of the agent at each moment between $t_1$ and $t_n$" (47), which seems quite right, even if we should resist or at least be skeptical about Tenenbaum's stronger suggestion/gloss that "an agent might be rational at each moment $t_x$ such that $t_x$ is within

[6] For in-depth discussion pertaining to this possibility, see (Andreou 2014).

the interval $t_{[1]}$ to $t_n$, and yet not be rational at interval $t_1$-$t_n$" (48).

Turn next to Tenenbaum's view that an instrumentally rational agent will often have to seek "acceptable" realizations of her ends rather than maximizing, and not due to bounded rationality but due to the nature of the ends themselves. Consider Warren Quinn's puzzle of the self-torturer,[7] which Tenenbaum describes (with some discretionary adjustments) as follows:

> A person has agreed to wear a device that delivers a constant but imperceptible electric shock. She, the self-torturer (ST), is then offered the following trade-off: she will receive a large sum of money – say, $100,000 – if she agrees to raise the voltage on the device by a marginal, that is, imperceptible or nearly imperceptible, amount. She knows that she will be offered this same trade-off again each time she agrees to raise the voltage. It seems that, at each step of the way, the agent should and would raise the voltage; after all, each rise in voltage makes at most a marginal difference in pain, well worth a gain of $100,000. But in so doing, she would eventually find herself in unbearable pain, and would gladly return all of the money, even pay some in addition, to be restored to the initial setting, at which she was poor but pain-free. Thus the ST appears to face a dilemma: no matter which choice she makes – continue indefinitely or stop at some point – her action seems irrational, or leads quickly to a state of affairs that no rational agent would accept: If she continues indefinitely she continually loses money for no gain, while if she stops she fails to act on her preferences. (85)

As Tenenbaum emphasizes, "the self-torturer has … two fairly ordinary ends (roughly avoiding pain and making money) … [that] generate a very clear (though not well-behaved) preference ordering" (83-84). More specifically, the self-torturer's preferences over the options are cyclic in that, for each pair of *adjacent* settings, the self-torturer prefers to stop at the higher setting rather than the lower setting and yet there is a sufficiently high setting n (among many sufficiently high settings) which is such that the self-torturer prefers stopping at the initial setting at which the voltage is not raised at all over stopping at setting n. Though "perfectly innocent from the point of view of instrumental rationality," the self-torturer's ends make maximizing with respect to the preferences they generate impossible (100). For every setting, there is an alternative setting that the self-torturer prefers. And yet, as Quinn suggests, and as Tenenbaum and I accept as common ground between us, we, as theorists of instrumental rationality, are being "too easy on [ourselves]" and "too hard on the self-torturer" if we simply dismiss the self-torturer's preferences as irrational (Quinn 1993: 199). Instrumental rationality must, it seems, prompt the agent to stop at an acceptable stopping point. This is an intriguing and tricky idea. How shall we understand the notion of acceptability?

[7] See (Quinn 1993).

Insofar as some stopping points are supposed to be acceptable and some are not, even though maximization is out of the question, the standard of acceptability cannot be that an option is acceptable only if there is no higher-ranked option. Tenenbaum suggests that an acceptable option is one that is "good enough" or satisfactory, but it seems like, even when maximizing is not possible, settling for an option that is "good enough" (from the agent's perspective) is misguided if, for example, options that are great (from the agent's perspective) are available.[8] Suppose, for example, that the self-torturer stops at setting 0, which she deems satisfactory, even though things would be great (from her perspective) were she to stop at setting 20 instead. This seems irrational. Why endorse the agent's stopping at a setting that qualifies (for her) as "satisfactory" (all-things-considered) when a setting that qualifies (for her) as "great" (all-things-considered) is available?

Significantly, my reasoning here draws on the distinction between *categorial subjective appraisal responses* and *relational subjective appraisal responses*.[9] Although this is not the place to delve into the distinction, the basic idea is as follows:

> [Loosely speaking,] relational subjective appraisal responses rank options in relation to one another; it is these appraisal responses that are captured by the agent's preferences … By contrast, categorial subjective appraisal responses place options in categories, such as, for example, "great" or "terrible."[10] (Andreou, in press)

Like relational subjective appraisal responses, categorial subjective appraisal responses can vary from agent to agent. A's categorial subjective appraisal responses might categorize option *x*, say, eating these pickled tomatoes, as "great" while B's categorial subjective appraisal responses categorize the option as "terrible."

It might be suggested that, even if, relative to the agent's all-things-considered evaluations, the options fall along a spectrum of vaguely bounded evaluative categories like "terrible," "bad," "satisfactory," "good," and "great," the fact that there is always an option that is preferred over any option the agent considers implies that there will always be an option that falls into a higher evaluative category than any option the agent considers, and so, like maximizing relative to the agent's preferences, seeking to settle on an option in the highest evaluative category in play is also impossible. But this does not follow. All the options in the case of the self-torturer might fall within a finite spectrum of categories

[8] Here and in the next few paragraphs, I draw on (Andreou 2015).

[9] See (Andreou, in press) and (Andreou 2015); the latter uses slightly different terminology.

[10] I here loosely describe "the favoring of one option in a pair as the ranking of that option over the other – even when no ranking of all the options is to be had because the agent's preferences are cyclic. If my use of 'ranking' seems too loose, it can be eliminated by the reader via appropriate substitutions" (Andreou, in press).

with, say, low settings falling somewhere in the ballpark of bad and/or satisfactory, mid-range settings falling somewhere in the ballpark of satisfactory, good, or great, and higher settings "circling back" through satisfactory and bad to terrible. But then, even though maximizing relative to the agent's preferences remains out of the question, settling for a satisfactory option seems rash.

I propose that we (partially) characterize (rational) acceptability as follows: An option is acceptable only if there is no higher-ranked option or, if there are no maximal options (where a maximal option is such that there is no higher-ranked option), only if the option falls squarely within the highest (all-things-considered) evaluative category in play.[11] (I here restrict my attention to cases where there is a finite number of ordered categories in play, as, tangential complications aside, we can assume is the case in the self-torturer's predicament.) Where there are no maximal options, as in the case of the self-torturer, settling on an option that is acceptable according to the preceding characterization will necessarily involve satisficing in the sense of settling on an option which is such that a higher-ranked option is available. It need not, however, involve settling on an option that is "good enough" or "satisfactory" in an intuitive sense. For instance, where there are no maximal options and the evaluative categories in play are, say, just "bad" and "terrible," ending up with a bad option will qualify as (rationally) acceptable even though it falls short of ending up with a(n) (intuitively) satisfactory option. Relatedly, where the evaluative categories in play are, say, just "satisfactory" and "good," ending up with a satisfactory (as contrasted with good) option will *not* qualify as (rationally) acceptable. Although my suggested proposal for understanding acceptability glosses over a large number of complications (which I broach elsewhere),[12] it is, I hope somewhat illuminating with respect to the intriguing but tricky idea that an instrumentally rational agent will often have to seek "acceptable" realizations of her ends rather than maximizing, and not due to bounded rationality but due to the nature of the ends themselves.

Turn finally to the idea that an agent may be rationally permitted to waver between options in a way that involves him incurring costs that he could have avoided if he resisted "brute shuffling," though not if the result is "disastrous" (156). Consider Tenenbaum's $200 WASTED case:

> Larry is deciding between being a professional footballer or a stay-at-home dad. In order to become a professional footballer, he must buy a $200 ball and net set. If he wants to be a stay-at-home dad, he needs to buy the *How to Be a Stay-at-Home Dad* DVD for $200. Larry forms the intention to become a professional footballer, goes to the store, and buys the ball and net set. Ten minutes later he abandons his intention,

[11] See (Andreou 2015).

[12] See (Andreou 2015).

calls the Barcelona manager, and says that he no longer wishes to be on the team as he is now a stay-at-home dad. (153)

Suppose this is a case in which Larry finds being a professional footballer and being a stay-at-home dad incommensurable (with the options being unrankable for Larry as one better than the other or as exactly equally good). (Like some assumptions flagged above, the assumption that two options can be incommensurable is controversial but one that Tenenbaum and I accept as common ground.) Suppose, relatedly, "that a difference of $200 dollars in the cost of either alternative would not suddenly make one of the options better than the other" (153). As such, Larry's choosing to be a stay-at-home dad would be permissible even if being a stay-at-home dad cost $200 more than originally anticipated. Still, as Tenenbaum grants, Larry's two choices (in the passage quoted above) seem collectively "foolish" – "it seems that something went awry" (154). Despite this appearance, Tenenbaum suggests that, other things equal, Larry does not count as irrational. Tenenbaum does grant that insofar as "repeated changes of mind would lead [Larry] to an unacceptable actualization of his pursuit of enough financial resources," Larry would, in a case involving such repeated changes of mind, qualify as irrational (156). Here, again, we run into the notion of an acceptable – in the sense of satisfactory or good enough – option. But, again, why settle on such acceptability? What if repeated changes of mind lead Larry from a good financial state to one that is, though not disastrous, only merely satisfactory? Hasn't something gone awry? The answer, I contend, is yes. Although satisfactory, the result is rationally unacceptable for essentially the same reason that it would be if the result were disastrous (and Tenenbaum grants that the result would be rationally unacceptable then) – the reason is that the agent failed to settle on an option in the highest evaluative category in play. Tenenbaum might be willing to grant this, maintaining that, when such failure is at issue, other things are not equal; they count as equal only when the waste in financial resources is small. But, assuming now that satisfactoriness is not enough for rational acceptability in cases of incommensurability, one might wonder why one should count as acceptable a series of choices that realize, over time, a non-maximal option. Unlike in cases involving (rationally innocent) cyclic preferences, realizing (through one's choices over time) a maximal option seems like something that a rational agent can and should aspire to in cases of incommensurability (given that, as I think both Tenenbaum and I assume, temporally extended agents are accountable [other things equal] for how their choices over time add up, and, in particular, for avoiding self-defeating patterns of choice). But then something *has* gone awry when one has proceeded in a way that is wasteful, even when one's choices over time generate only a small, rather

than disastrously large, waste of resources. The widely shared intuition that Larry has made a mistake should not, I think, be abandoned.

A reader less sympathetic than myself to Tenenbaum's general approach may balk at many of the assumptions that Tenenbaum and I both accept and that I incorporate into my discussion without defense. My aim has not been to defend our shared premises but, taking them as given, to advance our general approach further by complicating and enriching debate via the consideration of subtleties that merit further consideration and exploration. According to my reasoning, proceeding acceptably over time may well involve proceeding acceptably *at* each moment, even if, as Tenenbaum maintains, "the rationality of an agent through a time interval $t_1$ to $t_n$ does not supervene on the rationality of the agent at each moment between $t_1$ and $t_n$" (47). Moreover, proceeding acceptably over time, though it may, even apart from considerations of bounded rationality, often involve satisficing in the sense of settling on an option which is such that a higher-ranked option is available, this need not amount to settling on an option that is "good enough" or "satisfactory" in an intuitive sense. Instead, rationality may, if there are no maximal options, require one to settle on an option in the highest available evaluative category, which may be better or worse than "satisfactory." Finally, given the availability of one or more maximal options, as in cases of incommensurability, a rational agent can and should aspire to realize (through her choices over time) a maximal option, which requires avoiding "wasteful" instances of "brute shuffling," even if the result of such brute shuffling is "good enough."

Chrisoula Andreou
Department of Philosophy, University of Utah
c.andreou@utah.edu

## *References*

Andreou, C., 2014, "The Good, the Bad, and the Trivial," in *Philosophical Studies* 169: 209-225.

—, 2015, "The Real Puzzle of the Self-Torturer: Uncovering A New Dimension of Instrumental Rationality," in *Canadian Journal of Philosophy* 45: 562-575.

—, (in press), *Choosing Well*, Oxford University Press, New York.

Bratman, M., 2012, "Time, Rationality and Self-Governance," in *Philosophical Issues* 22: 73-88.

Quinn, W., 1993, "The Puzzle of the Self-Torturer," in *Morality and Action*, Cambridge University Press, Cambridge.

Tenenbaum, S., 2020, *Rational Powers in Action*, Oxford University Press, Oxford.

Thompson, M., 2008, "Naïve Action Theory," in *Life and Action*, Harvard University Press, Cambridge MA, 83-146.